A Method for Deriving Quasi-healthy Cohorts From Clinical Data

Biomedical Informatics Insights
Volume 10: 1–10
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1178222618777758

Satoshi Irino[1] and Yukio Kurihara[2]

[1]Department of Nursing, Ehime Prefectural University of Health Sciences, Tobe-cho, Japan.
[2]Nursing Course of Kochi Medical School, Nankoku, Japan.

ABSTRACT: We evaluated quasi-healthy cohorts (model cohorts), derived from clinical data, to determine how well they simulated control cohorts. Control cohorts comprised individuals extracted from a public checkup database in Japan, under the condition that their values for 3 basic laboratory tests fall within specific reference ranges (3Ts condition). Model cohorts comprised outpatients, extracted from a clinical database at a hospital, under the 3Ts condition or under the condition that their values for 4 laboratory tests fall within specific reference ranges (4Ts condition). Because even a patient with a serious illness, such as cancer, may present with normal values on basic laboratory tests, one additional condition was added: the duration (1 or 3 months; 1M or 3M) during which patients were not hospitalized after their first laboratory test. For evaluations, cohorts were specified by age and sex. The 4Ts + 3M condition was the most effective condition, under which model cohorts were used to successfully simulate age-dependent changes and sex differences in laboratory test values for control cohorts. Therefore, by properly setting the conditions for extracting quasi-healthy individuals, we can derive cohorts from clinical data to simulate various types of cohorts. Although some issues with the proposed method remain to be solved, this approach presents new possibilities for using clinical data for cohort studies.

KEYWORDS: Cohort study, secondary use, clinical data, statistical analysis, derivation of cohorts

RECEIVED: September 15, 2017. **ACCEPTED:** April 25, 2018.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Satoshi Irino, Department of Nursing, Ehime Prefectural University of Health Sciences, 543 Takooda, Iyo-gun, Tobe-cho 791-2101, Ehime, Japan. Email: sirino@epu.ac.jp

Background and Significance

By the mid-1990s, the computerized physician order entry (CPOE) system was being widely used,[1] and enormous amounts of clinical data were being collected at large hospitals. Secondary uses of this vast quantity of clinical data have been an active area of research since the beginning of 2000. Clinical data usually comprise data of individuals with diseases and have been used for assessing clinical treatments, understanding illness progression, and finding new diagnostic methods.[2–6] Searching the contents of clinical data from the present to the past is, therefore, the current mainstream method of using such data. Paradoxically, if a healthy cohort can be established, under certain conditions, in a clinical database, the process of disease onset from the past to the present can be verified, as is done in cohort studies.[7] This notion was the basis for our research.

Since the 1960s, clinical data have been used in laboratory medicine for deriving the normal or reference ranges for clinical laboratory tests.[8–10] This use was based on the idea that the excluding values for clinical data, which are statistically too low or too high, may be statistically similar to the data obtained from healthy individuals.[11–13] This suggests that it is possible to extract individuals at various levels of health from clinical data by changing the extraction conditions. Herein, we refer to individuals who have minor health issues but are otherwise healthy, as quasi-healthy individuals. If, using numerous extraction conditions, we extract from the clinical data individuals whose values for multiple tests are within the reference ranges, the level of health for these individuals can be assumed to be very high. Conversely, if we extract individuals using fewer extraction conditions, their level of health must be low. Therefore, using quasi-healthy individuals, we can derive a cohort to simulate the cohort of interest, and we could conduct various cohort studies less expensively.

Residents of Japan, who have a health insurance card, are not required to use a specific health care provider and can directly visit any hospital, including a university hospital.[14,15] Consequently, the number of outpatients at hospitals in Japan is 3 to 4 times greater than that in foreign countries.[16] Therefore, among these outpatients, there may be numerous quasi-healthy individuals, even at university hospitals. University hospitals are advantageous for data studies because these institutions generally accumulate clinical data over a long period and have high quality control.

Between 1983 and 2007 in Japan, each local government encouraged all inhabitants 40 years of age and older to receive an annual public checkup. A new public checkup system, which targets the metabolic syndrome, was initiated in 2008, but the types of tests remain limited.[17] The inhabitants can participate in the public checkup every year. The results of these checkups have been accumulated by each local government. Because approximately 70% of these public checkup examinees did not require detailed testing,[18] ie, they did not have a serious illness, we were able to use those data to evaluate quasi-healthy individuals extracted from clinical data.

In this study, we propose a method to derive cohorts of quasi-healthy individuals from clinical data. By properly

Table 1. The number of checkup examinees and ratio among the population of Kochi Prefecture from the period of 2003 to 2007.

AGE BRACKETS, Y	40-49	50-59	60-69	70-79	ALL AGE BRACKETS
Male	7579 (3%)	15306 (5%)	31947 (13%)	35081 (17%)	89914 (9%)
Female	16471 (7%)	37033 (12%)	61440 (22%)	58104 (21%)	173047 (16%)
Males + females	24050 (5%)	52339 (9%)	93387 (18%)	93185 (19%)	262961 (13%)

setting conditions to extract quasi-healthy individuals, we show how well the cohorts of quasi-healthy individuals can simulate the statistical features of cohorts of healthy individuals derived from a database of public checkup examinees.

In section "Subjects and Methods," we describe the extraction of control subjects from the public checkup data of the Kochi Prefecture and derivation of control cohorts. We also describe the selection of quasi-healthy subjects from the clinical data at Kochi Medical School Hospital (KMSH) under several health-level conditions and the derivation of quasi-healthy model cohorts (hereafter, model cohorts) used to simulate control cohorts. In section "Results," we show the results of studies on statistical similarity between the control and model cohorts for different health-level conditions. In section "Discussion," we discuss the results, and in section "Conclusion," we present our conclusions.

Subjects and Methods

Data resources

The KMSH, located in western Japan, is one of the main general hospitals in the Kochi Prefecture. The KMSH has been using the CPOE system, called the Integrated Medical Information System, since October 1981.[19,20] Clinical data from approximately 290000 patients registered in this system have been available as of 2012. The number of patients visiting KMSH for annual medical consultations equals approximately 10% of the population of Kochi Prefecture. The clinical data were coded for epidemiologic studies. We extracted those coded data for this study under the approval of the Ethics Review Board of the Kochi Medical School. We designated this data set as the Clinical DS. The Kochi Health Research Institute has accumulated the results of public checkups performed in the Kochi Prefecture from 1989 to 2010. We obtained permission to analyze the coded data from the Kochi Health Research Institute and designated this data set as the Public Checkup DS.

To compare the statistical features of clinical laboratory test data in the Clinical DS with those in the Public Checkup DS, there must be equivalent levels of quality control between the 2 data sets. In particular, the Public Checkup DS consists of clinical test data from numerous laboratories; therefore, similar quality control levels among the laboratories are very important if the data are to be compared. In Japan, the quality control levels among laboratories were gradually standardized after 2001.[21] As mentioned in the previous section, a new public checkup system was initiated in 2008. Therefore, we compared data in the Clinical DS with those in the Public Checkup DS from 2003 to 2007. Table 1 shows, by sex and age bracket, the number of checkup examinees and ratio among the population in the Kochi Prefecture for the period of 2003 to 2007. On average, approximately 53000 inhabitants (18000 men and 35000 women) had annual public checkups during this period.

Deriving control and model cohorts

We used the Public Checkup DS to derive the control cohorts. Because the health level of public checkup examinees can vary substantially, the cohorts of examinees were not directly used as healthy cohorts or control cohorts. Instead, the control cohorts comprised individuals extracted from the Public Checkup DS whose values for the 3 tests were within the reference ranges. The 3 tests were blood hemoglobin (Hb) as the anemia index, blood creatinine level (CRN) as the renal function index, and alanine aminotransferase (ALT) as the liver function index. These 3 tests were used as indicators of whether the examinees had to undergo a more detailed examination. As shown in the "Results" section, the values of 75% to 80% of the examinees for the 3 tests fell within the reference ranges; this is consistent with the percentages reported in a previous section.[18] These tests were conducted on the same day to evaluate the basic health condition of the examinee. We designated this set of 3 tests as the 3Ts condition.

To derive the model cohorts from the patients registered in the Clinical DS, we set several conditions for the selection of patients. The procedure for deriving the cohorts is shown in Figure 1. First, we limited the patients to only those who sought outpatient medical consultations for the first time. This minimizes the effects of treatment, such as those of surgery performed in a hospital or those of medication. Second, patients needed to meet the 3Ts condition. Public checkup examinees typically do not have severe pathological symptoms such as a high fever; however, outpatients usually present with prevailing health issues, which is likely their reason for seeking a medical consultation. Therefore, additional indexes may be needed when deriving model cohorts from the Clinical DS. Here, we derived the model cohorts by adding the white blood cell (WBC) count as the infection index. We required the patients to have WBC counts within the reference ranges. We designated this set of 4 tests as the 4Ts condition. We used the

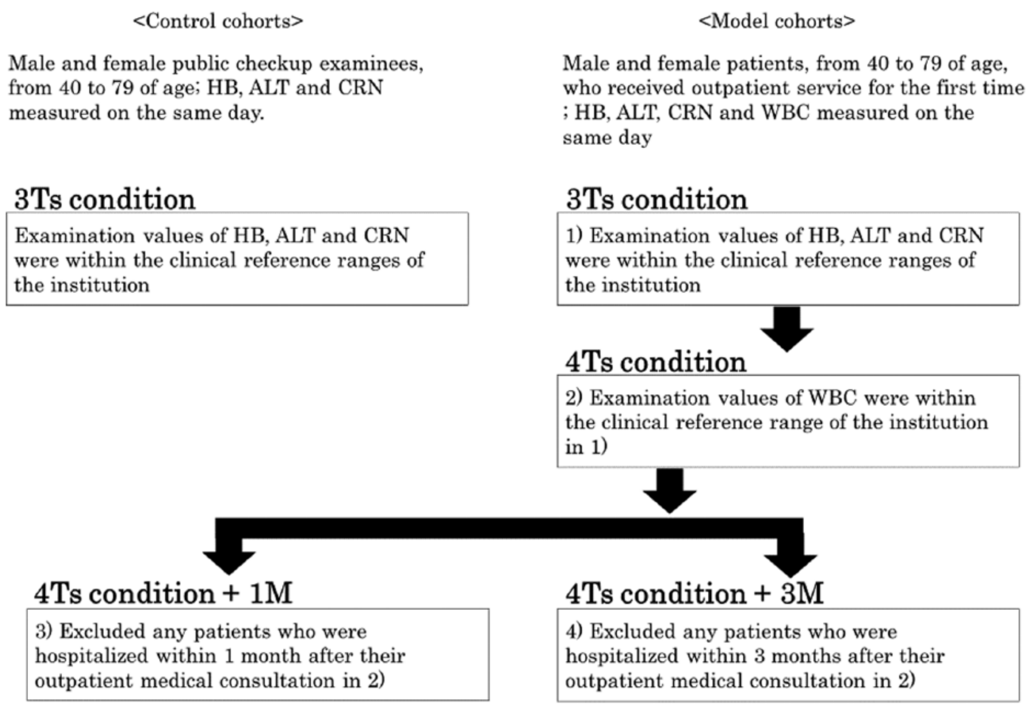
<Control cohorts>
Male and female public checkup examinees, from 40 to 79 of age; HB, ALT and CRN measured on the same day.
3Ts condition
Examination values of HB, ALT and CRN were within the clinical reference ranges of the institution
<Model cohorts>
Male and female patients, from 40 to 79 of age, who received outpatient service for the first time ; HB, ALT, CRN and WBC measured on the same day
3Ts condition
1) Examination values of HB, ALT and CRN were within the clinical reference ranges of the institution
4Ts condition
2) Examination values of WBC were within the clinical reference range of the institution in 1)
4Ts condition + 1M
3) Excluded any patients who were hospitalized within 1 month after their outpatient medical consultation in 2)
4Ts condition + 3M
4) Excluded any patients who were hospitalized within 3 months after their outpatient medical consultation in 2)

Figure 1. Procedure for extracting the cohorts.

Table 2. Clinical reference ranges used at the Kochi Medical School Hospital.

TESTS	MALES		FEMALES	
	LOWER LIMIT	UPPER LIMIT	LOWER LIMIT	UPPER LIMIT
Hb, g/dL	13.2	17.2	10.8	14.9
ALT, IU/dL	8	42	6	27
CRN, mg/dL	0.6	1.1	0.4	0.8
WBC, cells/μL	3600	9600	3000	8500

Abbreviations: ALT, alanine aminotransferase; CRN, blood creatinine level; Hb, hemoglobin; WBC, white blood cell.

clinical reference ranges, used at KMSH, as the reference ranges for this study (see Table 2).

Even a patient with a serious illness, such as cancer, may present with normal values on basic laboratory tests. Therefore, in addition to the 3Ts and 4Ts conditions, we also considered the duration over which patients were not hospitalized after their first laboratory test at an outpatient department. In Japan, there is a high referral rate of patients from other medical institutions to medical school hospitals. The referral rate for KMSH is approximately 60%. Therefore, some patients seeking outpatient medical consultations at medical school hospitals are hospitalized immediately, and some patients are followed up for several months. We considered 1 month (1M) and 3 months (3M) after having their first laboratory test as the periods until a patient with a severe illness or a high-risk patient was admitted to the hospital. As an additional condition to derive model cohorts from patients, we required model cohort to not have been admitted within 1M or 3M after having their first laboratory test; we designated these conditions as 4Ts + 1M or 4Ts + 3M, respectively.

We then specified the cohorts by sex and age. As we mentioned previously, local governments in Japan routinely perform public checkups for inhabitants 40 years of age and older, and the average life spans for Japanese men and women are 80 and 86 years, respectively.[8] Therefore, we set the ages of subjects from 40 to 79 years and divided the data into 4 age brackets with ranges of 10 years: 40-49, 50-59, 60-69, and 70-79.

Comparison of the model and control cohorts

Comparison of frequency distributions of laboratory test values. We examined the laboratory tests used and not used for deriving the cohorts. Although the values of the tests, used for deriving the cohorts, are limited, the similarities in frequency distributions of these values, in both model and

Table 3. The number and percentages of public checkup examinees extracted for the control cohorts and of the patients used for the model cohorts.

AGE BRACKETS, Y		40-49	50-59	60-69	70-79	ALL AGE BRACKETS
Control cohorts						
Males		5684 (75%)	12 245 (80%)	24 919 (78%)	24 206 (69%)	67 054 (75%)
Females		12 847 (78%)	29 997 (81%)	49 766 (81%)	45 902 (79%)	138 512 (80%)
Males + females		18 531 (77%)	42 242 (81%)	74 685 (80%)	70 108 (75%)	205 566 (78%)
Model cohorts						
Males	3Ts	346 (53%)	685 (55%)	815 (52%)	722 (43%)	2568 (50%)
	4Ts	301 (46%)	600 (48%)	740 (47%)	669 (40%)	2310 (45%)
	4Ts + 1M	172 (26%)	291 (23%)	310 (20%)	232 (14%)	1005 (20%)
	4Ts + 3M	146 (22%)	232 (19%)	233 (15%)	174 (10%)	785 (15%)
Females	3Ts	593 (65%)	913 (65%)	978 (63%)	1160 (62%)	3644 (63%)
	4Ts	526 (58%)	833 (60%)	914 (59%)	1074 (58%)	3347 (58%)
	4Ts + 1M	299 (33%)	479 (34%)	467 (30%)	419 (22%)	1664 (29%)
	4Ts + 3M	240 (26%)	395 (28%)	390 (25%)	342 (18%)	1367 (24%)
Males + females	4Ts + 3M	386 (25%)	627 (24%)	623 (20%)	516 (15%)	2152 (20%)

control cohorts, are not necessarily apparent. Thus, we first compared the frequency distributions of the laboratory test values used for deriving the cohorts; this was done to investigate the differences between the 2 cohorts. In this study, we examined Hb and ALT. The lower limit of CRN was set in the Public Checkup DS, but the frequency distributions of CRN were distorted; therefore, we did not examine this factor. We compared the frequency distributions of the laboratory test values of the cohorts using the Mann-Whitney (M-W) and Kolmogorov-Smirnov (K-S) tests (IBM SPSS Statistics Base, version 22.0). In this study, the "absence" of a significant difference in the results of these tests is important.

Next, we compared the frequency distributions of the test values not used for deriving the cohorts. We showed how well the differences between the frequency distributions of the control cohort and those of the model cohort, can be diminished by adding the health-level conditions. Here, we examined the levels of serum total cholesterol (TC) and γ-glutamyltransferase (γ-GTP), both of which had the largest amount of available data in the 2 data sets.

Evaluation of percentile errors by bootstrap method. We could not extract a sufficient number of individuals from the Clinical DS for the model cohorts; therefore, the statistical quantities calculated from the model cohort may be prone to error. Thus, we used the bootstrap method (IBM SPSS Statistics Bootstrapping, version 20.0) to evaluate the statistical errors of the statistical quantities calculated from the distribution of the test values. Because the agreement of the central parts of the distributions is critical, we evaluated the statistical errors of the 25th, 50th, and 75th percentiles in both cohorts and checked whether the 2 ranges of error overlapped.

Results

The numbers and average ages of individuals used for control and model cohorts

Table 3 shows the number of individuals used for the control and model cohorts under the conditions 3Ts, 4Ts, 4Ts + 1M, and 4Ts + 3M specified by the sex and age bracket.

For the control cohorts, the values in parentheses show the extraction percentages of individuals from the control cohort, as related to all the inhabitants who were tested for Hb, CRN, and ALT on the same day. The extraction percentages for males and females were 75% to 81 % in the 40-69 age bracket. However, the extraction percentage for males in the 70-79 age bracket was 69%, which was approximately 10% lower than those in the other age brackets. The extraction percentage for females in the 70-79 age bracket was similar to that for females in the other age brackets.

For the model cohorts, the values in parentheses show the extraction percentages of individuals from the model cohorts, as related to all patients who were tested for Hb, CRN, ALT, and WBC for the first time on the same day at the outpatient departments of KMSH. For males, the extraction percentage under the condition of 3Ts was 53% in the 40-69 age bracket. The

Table 4. The average ages for the control and model cohorts under the 4Ts + 3M condition.

AGE BRACKETS, Y	MALES				FEMALES			
	CONTROL COHORTS		MODEL COHORTS (4TS + 3M)		CONTROL COHORTS		MODEL COHORTS (4TS + 3M)	
	AVERAGE ± SD	MEDIAN	AVERAGE ± SD	MEDIAN	AVERAGE ± SD	MEDIAN	AVERAGE ± SD	MEDIAN
40-49	45.0 ± 2.9	45.0	44.5 ± 2.9	45.0	44.8 ± 2.9	45.0	44.9 ± 2.9	45.0
50-59	55.1 ± 2.8	55.0	54.8 ± 2.8	55.0	55.2 ± 2.7	55.0	54.7 ± 2.7	55.0
60-69	65.0 ± 2.8	65.0	64.2 ± 2.9	64.0	64.7 ± 2.8	65.0	64.3 ± 2.8	65.0
70-79	73.8 ± 2.8	74.0	73.2 ± 2.8	73.0	73.9 ± 2.8	74.0	74.4 ± 2.8	74.0

extraction percentage under the 4Ts + 3M condition was 19% in the 40to 69 age bracket and 10% in the 70-79 age bracket. For females, the extraction percentages under the condition of 3Ts were 64% for all of the age brackets. The extraction percentage under the condition of 4Ts + 3M was approximately 26% in the 40-69 age bracket and 18% in the 70-79 age bracket.

Table 4 shows the average ages, standard deviation, and median for the control and model cohorts under the condition of 4Ts + 3M. The differences in the average ages for both cohorts were within 0.5 years for every age bracket. Moreover, the standard deviation and median for each cohort were similar.

Comparison of the frequency distributions of laboratory tests used for deriving cohorts

The *P* values of the M-W and K-S tests for Hb and ALT are shown in Table 5.

For males in the 40-49 age bracket, the M-W and K-S tests did not show any significant differences under the conditions of 3Ts, 4Ts, 4Ts + 1M, and 4Ts + 3M. For females in the 40-49 age bracket, the M-W test showed a significant difference in Hb under the condition of 4Ts + 3M.

For males in the 50-59, 60-69, and 70-79 age brackets, the K-S test, the M-W test, or both tests showed significant differences in Hb under the condition of 3Ts; however, significant differences were not observed under the condition of 4Ts + 3M. In the 60-69 age bracket, the M-W and K-S tests showed significant differences in ALT under the 3Ts condition. Significant differences in ALT, as assessed by the M-W test, were observed under the condition of 4Ts + 3M.

For females in the 50-59 age bracket, the K-S test showed significant differences in Hb under the 3Ts condition, but not under the 4Ts + 3M condition. In the 60-69 age bracket, the M-W and K-S tests showed significant differences in ALT under the 3Ts condition; however, significant differences in ALT, as assessed by the M-W test, were not observed under the 4Ts + 3M condition. In the 70-79 age bracket, the M-W and K-S tests showed significant differences in ALT under the 3Ts condition, but not under the 4Ts + 3M condition.

There were no significant differences in Hb and ALT between the model and control cohorts under the 4Ts + 3M condition; the exceptions were the levels of ALT in males in the 60-69 age bracket and Hb in females in the 40-49 age bracket.

Comparison of frequency distributions of the laboratory tests not used for deriving cohorts

We compared the frequency distributions of the laboratory tests not used for deriving cohorts. The *P* values of the M-W and K-S tests for TC and γ-GTP are shown in Table 5.

For both males and females in the 40-49 age bracket, the M-W and K-S tests did not show significant differences in values under the 3Ts, 4Ts, 4Ts + 1M, and 4Ts + 3M conditions. In the 50-59 age bracket, the M-W and K-S tests showed significant differences in TC under the 3Ts condition, but not under the 4Ts + 3M condition. In the 60-69 and 70-79 age brackets, the M-W and K-S tests showed significant differences in TC and γ-GTP under the 3Ts condition. Significant differences in TC and γ-GTP in the 60-69 age bracket were not observed under the 4Ts+3M condition; however, significant differences in γ-GTP, as assessed by the K-S test, were observed for males in this age bracket. The significant differences in TC and γ-GTP in the 70-79 age bracket were not observed under the 4Ts + 3M condition.

Most significant differences in the values, obtained by both tests, were not observed under the 4Ts + 3M condition; however, significant differences in γ-GTP, as assessed by the K-S test, were observed for males in the 60-69 age bracket.

Comparison of percentile errors

We used the bootstrap method to evaluate the statistical errors for the 25th, 50th, and 75th percentiles of both control and model cohorts, for males and females, and for all the performed laboratory tests (see Figures 2 and 3). The Public Checkup DS contains a large amount of data; therefore, the statistical errors were very small.

The sizes of the errors for Hb, ALT, TC, and γ-GTP in the 25th, 50th, and 75th percentiles for both males and females in the 60-69 age bracket were greater than those in the other age brackets. Among the 4 tests, the fluctuation in the sizes of the errors for ALT and γ-GTP in males was relatively great.

Table 5. The *P* values for the Mann-Whitney and Kolmogorov-Smirnov tests.

SEX	AGE BRACKETS, Y	CONDITIONS	MANN-WHITNEY TEST				KOLMOGOROV-SMIRNOV TEST			
			HB	ALT	TC	Γ-GTP	HB	ALT	TC	Γ-GTP
Males	40-49	3Ts	.875	.260	.242	.603	.946	.653	.363	.911
		4Ts	.781	.388	.153	.503	.985	.840	.207	.574
		4Ts + 1M	.365	.725	.425	.139	.716	.796	.600	.282
		4Ts + 3M	.361	.330	.157	.122	.543	.647	.192	.321
	50-59	3Ts	.028*	.738	<.001**	.652	.015*	.520	.003**	.709
		4Ts	.004**	.607	.001**	.988	.006*	.667	.010*	.800
		4Ts + 1M	.287	.990	.214	.623	.724	.637	.445	.609
		4Ts + 3M	.178	.917	.160	.923	.401	.801	.258	.869
	60-69	3Ts	.002**	.007*	.020*	<.001**	.019*	.003**	.015*	<.001**
		4Ts	.012*	.004**	.036*	<.001**	.030*	.003**	.033*	<.001**
		4Ts + 1M	.886	.001**	.420	.002**	.836	.038*	.648	.003**
		4Ts + 3M	.866	.023*	.494	.054	.868	.284	.651	.042*
	70-79	3Ts	.028*	.299	.006**	.007**	.095	.099	.005**	.023*
		4Ts	.047*	.401	.012*	.008**	.146	.157	.023*	.037*
		4Ts + 1M	.210	.956	.303	.267	.511	.995	.589	.101
		4Ts + 3M	.487	.917	.144	.141	.956	.997	.397	.111
Females	40-49	3Ts	.188	.252	.557	.427	.189	.255	.144	.228
		4Ts	.436	.424	.575	.963	.405	.594	.291	.431
		4Ts + 1M	.122	.647	.297	.956	.206	.767	.050	.666
		4Ts + 3M	.027*	.594	.268	.667	.070	.636	.077	.415
	50-59	3Ts	.464	.001**	.002**	.123	.026*	.004**	<.001**	.227
		4Ts	.857	.005**	.003**	.172	.100	.016*	.001**	.377
		4Ts + 1M	.942	.344	.450	.308	.45	.435	.452	.575
		4Ts + 3M	.659	.630	.382	.668	.307	.675	.300	.913
	60-69	3Ts	.841	.033*	.005**	.001**	.327	.001**	.004**	.002**
		4Ts	.778	.077	.053	.014*	.539	.004**	.037*	.050
		4Ts + 1M	.041*	.265	.153	.035*	.114	.030*	.083	.194
		4Ts + 3M	.205	.246	.246	.267	.366	.074	.177	.673
	70-79	3Ts	.150	<.001**	<.001**	.019*	.078	<.001**	<.001**	.043*
		4Ts	.159	.001**	.006**	.082	.107	<.001**	.015*	.088
		4Ts + 1M	.181	.169	.251	.264	.229	.118	.205	.092
		4Ts + 3M	.111	.240	.707	.240	.201	.195	.653	.153

Abbreviations: γ-GTP, γ-glutamyltransferase; ALT, alanine aminotransferase; Hb, hemoglobin; TC, total cholesterol.
*$P < .05$; **$P < .01$.
The test item with a significant difference was shown by an italic character.

The values of the 4 tests for the 25th, 50th, and 75th percentiles for males were nearly unchanged or increased and decreased gradually from the 40-49 age bracket to the 60-69 age bracket. In the 70-79 age bracket, the values of the 4 tests for the 25th, 50th, and 75th percentiles were lower than those in the other age brackets. Figure 2 shows that all of the

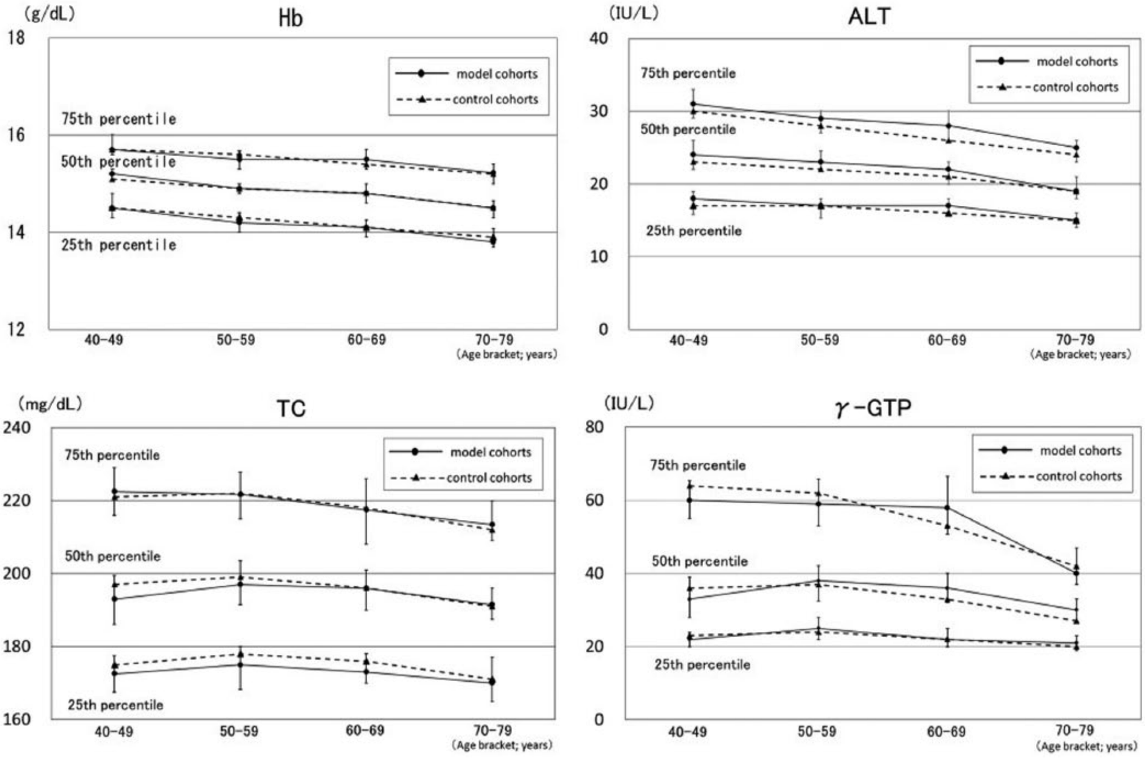

Figure 2. Evaluation of statistical errors in the 25th, 50th, and 75th percentiles, for males in the control and model cohorts, by the bootstrap method.

laboratory test values for males decreased gradually from middle to old age.

The values of the 4 tests for the 25th, 50th, and 75th percentiles for females increased from the 40-49 age bracket to the 50-59 age bracket. However, in the 50-59 age bracket to the 60-69 age bracket, the values of the 4 tests for the 25th, 50th, and 75th percentiles were nearly equal. These values then decreased in the 70-79 age bracket. These tendencies were evident in the values for TC. Figure 3 shows that all of the laboratory test values for females increased from the 40-49 age bracket to the 50-59 age bracket.

Discussion

Validity of the method for deriving the quasi-healthy model cohort

There are 3 key points in our discussion on the validity of this method, used to derive quasi-healthy cohorts (model cohorts). The first point is the validity of using a control cohort as healthy cohort, the second point is the validity of the conditions to derive a model cohort from clinical data, and the third point is the sufficiency of the comparison between a model cohort and a control cohort.

Although nearly 70% of the public checkup examinees did not require more detailed testing, this suggests that the remaining 30% of the examinees did require more detailed testing; this may indicate that their health level is low. We have to exclude as many of these individuals as possible when deriving control cohorts from the public checkup examinees. As shown in Table 3, we excluded 20% to 31 % of the male examinees and 19% to 22 % of the female examinees using the 3Ts condition. Therefore, the percentage of examinees with low health, included in the control cohorts, may be less than 10%. In this sense, the present control cohorts may properly represent cohorts of healthy individuals. We also considered the overlap between a cohort of outpatients visiting KMSH and a control cohort. Based on data shown in Tables 1 and 3, the ratios of male and female examinees included in a control cohort, to those in the population of the Kochi Prefecture, were .023 (.03 × .75) in the 40-49 age bracket to .117 (.17 × .69) in the 70-79 age bracket, and .054 (.07 × .77) in the 40-49 age bracket to .178 (.22 × .81) in the 60-69 age bracket, respectively. This means that the percentages of healthy male and female examinees, included among the outpatients, are maximally 11.7% in the 70-79 age bracket and 17.8% in the 60-69 age bracket, respectively. However, healthy examinees usually do not visit hospitals, and those percentages should be largely reduced. Consequently, the overlap between a cohort of outpatients visiting KMSH and a control cohort must be small.

Next, we discuss the validity of the conditions used to derive a model cohort: 3Ts, 4Ts, 4Ts + 1M, and 4Ts + 3M. As shown in Table 5, there were no significant differences between a model cohort extracted using the 3Ts condition and a control cohort in the 40-49 age bracket for both sexes. When younger patients have definitive health issues, such as infection or fracture, they visit a hospital; therefore, the basic 3Ts condition is very effective in this case. In contrast, because many older patients have chronic diseases, it is difficult to extract quasi-healthy

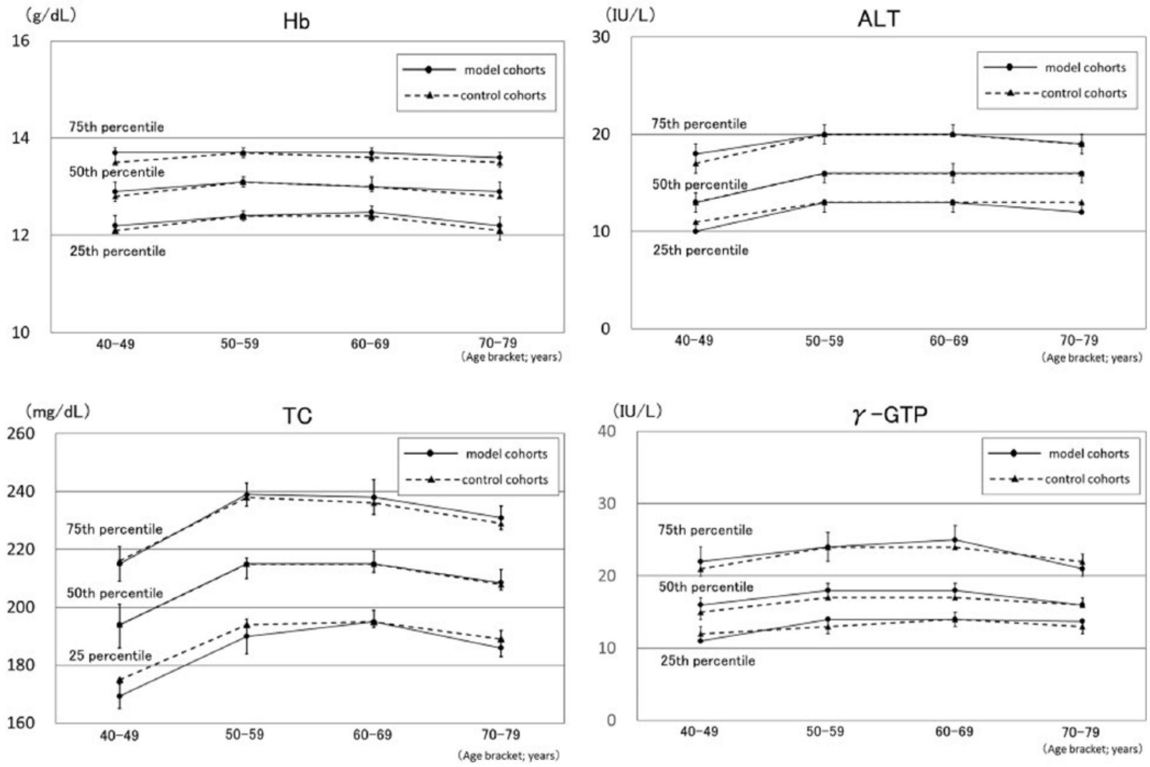

Figure 3. Evaluation of statistical errors in the 25th, 50th, and 75th percentiles, for females in the control and model cohorts, by the bootstrap method.

individuals with only a few tests. The 4Ts condition, set by adding the WBC test to the 3Ts condition, removed 6 significant differences between a model cohort extracted using the 3Ts condition and a control cohort in 3 age brackets for females. Adding the duration of no hospitalization after the first laboratory test to the 4Ts condition drastically reduced the number of significant differences between a model cohort, extracted using the 4Ts condition, and a control cohort. Because the patients who were admitted soon after their first laboratory test at an outpatient department likely had a serious problem, this result is reasonable. The condition adding the duration of no hospitalization is unique and is an important advantage when using clinical data.

Finally, we discuss the sufficiency of the comparison between a model cohort and a control cohort from 2 perspectives. First, we consider the significant differences in the comparison. We evaluated whether the differences between the frequency distributions of the values for the 4 laboratory tests (Hb, ALT, TC, and γ-GTP), for the model and control cohorts, were significant ($P < .05$) or not significant, as assessed by the M-W and K-S tests. A P value of $\geqslant .05$ does not mean that the 2 distributions are equivalent, however. In this sense, we only showed that there were no significant differences between the frequency distributions of the 4 laboratory tests for a model cohort satisfied under the 4Ts + 3M condition and a control cohort. However, as shown in Table 5, many P values for 4Ts + 3M are greater than .5. If the P value is greater than .5, the probability that the 2 distributions are equivalent is higher than that stating that they are different. In this case, we consider the model cohorts under the 4Ts + 3M condition to be good substitutions for the healthy cohorts.

Next, we consider the sex and age dependence of the laboratory tests. As shown in Figures 2 and 3, the model cohorts under the 4Ts + 3M condition acceptably simulate the age-related trends and differences between the sexes for the 4 laboratory tests. In particular, the large increase in TC, associated with menopause in females,[22–24] is well simulated by the model cohorts. When we consider age dependence of the laboratory tests, the age distribution in each age bracket is very important. As shown in Table 4, there were few differences in ages between the model and control cohorts. These results indicate good age distributions in this study. Therefore, the comparison between a model cohort and a control cohort is sufficient for examining laboratory tests.

Discrepancies in ALT, γ-GTP, and Hb

We evaluated the causes of statistical inconsistencies in ALT and γ-GTP for males in the 60-69 age bracket and in Hb for females in the 40-49 age bracket.

As shown in Figure 2, the values for the 75th percentile for γ-GTP in the 40-49, 50-59, and 60-69 age brackets, in both the model and control cohorts, were beyond the upper value of the reference range, ie, 50 IU/L. Alcohol consumption increases the levels of γ-GTP. Alcohol consumption per person in the study area of Kochi Prefecture is the second highest nationwide[25]; this may be associated with the high values for γ-GTP. The adverse effects of excessive alcohol consumption on health begin to appear approximately at the age of 50 years. Because the value of γ-GTP is correlated with that of ALT, the significant difference in ALT remained in the 60-69 age bracket; however, the values of ALT were within the reference range. Therefore, γ-GTP should be added to the extraction conditions to remove the effects of alcohol consumption. We could not add γ-GTP to the extraction conditions in this study because we did not have a sufficient number of subjects or indicators.

For females in the 40-49 age bracket, the frequency distribution of Hb for the model cohort, under all the extraction conditions except for 4Ts + 3M, did not significantly differ from that of the control cohort (Table 5). However, the values of the 3 percentiles for Hb in the model cohort, under the condition of 4Ts + 3M, were higher than those for the control cohort; the errors obtained from the bootstrap evaluation were relatively large, as shown in Figure 3. As shown in Table 3, the number of individuals in the model cohort, under the condition of 4Ts + 3M, was the smallest. The fluctuation of the test values in a small cohort is usually greater than that in a large cohort. Individuals with higher values of Hb may have accidentally remained in the model cohort. This problem can be resolved by increasing the number of samples, as mentioned in the next section.

Our evaluation of the 3 inconsistent items suggests that it is important to not only pay heed to any decrease in the number of subjects but also to have an understanding of the regional characteristics of the Clinical DS when using this method.

Possibilities of quasi-healthy model cohorts

The success of this study highlights the possibilities of using laboratory test data from a clinical database. We showed that a model cohort, simulating a certain cohort, can be derived from clinical data by properly setting the extraction conditions. Consequently, we may be able to derive various types of cohorts from clinical data by adjusting the extraction conditions based on data from quasi-healthy individuals. Previous studies on deriving normal or reference ranges for clinical laboratory tests only derived cohorts of healthy individuals from clinical databases.

The model cohorts did not significantly differ from the control cohorts derived from the Public Checkup DS, even when assessed by TC and γ-GTP tests, which were not used as extraction conditions. In the future, if no significant differences can also be confirmed for other tests, our method can enable using a model cohort in clinical and epidemiologic studies. An advantage of using clinical data is that they have more test items than the Public Checkup DS. For example, if test items, included in clinical data but not in the Public Checkup DS, can be evaluated and health problems can be identified, health measures can be designed more effectively. In this study, TC showed no inconsistencies throughout all age groups in both males and females; applications of this method already show promise in other lipid-based tests correlated with TC, such as high-density lipoprotein cholesterol and low-density lipoprotein cholesterol.

In future cohort studies, we may be able to predict trends in results by evaluating study designs in a preliminary investigation using clinical data. For example, we may be able to estimate trends in treatment courses and periods up until subjects in model cohorts are diagnosed with ischemic heart disease. However, we did not use any index of circulatory function as an extraction condition. Therefore, the quasi-healthy cohorts in this study may include subjects with circulatory diseases. When deriving a cohort with specific characteristics of interest, we will need to use an index related to the study purpose for the extraction condition. This study only shows one pattern for the extraction method, and it is important to modify the extraction conditions in accordance with studies to be conducted. Future cohort studies, using various types of model cohorts, may help refine the method for deriving model cohorts.

This study also revealed a problem with the high rate of decrease in the number of samples, with respect to the original data sample, when deriving model cohorts from clinical data. Securing a sufficient number of samples by linkage of clinical data between facilities is a possible solution to this problem; this may also help correct the difference in age distribution. Furthermore, dropouts are highly likely in follow-ups of the same cohort, derived from the clinical data of a single facility; this is due to the characteristics of the outpatient examinees. For example, some patients may be transferred to another hospital. However, patients with mild symptoms tend to continue visiting the same local medical institution, indicating that this method can currently be applied for certain types of cohort studies. In the future, we will conduct additional studies to verify the rate of continuation and continuation periods within the same database of clinical data and to resolve issues via clinical data linkages between multiple facilities.

Study limitations and further development

This study has some limitations. We used clinical data from one facility; therefore, the number of samples was limited. The number of subjects in the final derived model cohorts ranged from 146 to 233 for males and from 240 to 395 for females. The size of these model cohorts may not be sufficient for detailed comparisons. Moreover, we evaluated statistical features for only 4 laboratory tests; among these laboratory tests, only 2 tests were used for the extraction conditions. Consequently, only a few indicators were used for evaluations. Regarding the Clinical DS at the KMSH, used in this study, the patient referral rate from other medical institutions accounted for approximately 60% of the total patients. In other words, many patients, who might be in relatively poor health, were included in this Clinical DS. Therefore, we derived small cohorts from this Clinical DS to simulate healthy cohorts.

A possible solution to the problem of sample size is linkage of data sets among medical facilities. However, because additional problems may be encountered when linking data sets, such as different reference ranges among facilities and countries, more work will be required to better understand how to best derive model cohorts. In the future, an improved method for deriving model cohorts will contribute greatly to the overall goal of using vast quantities of clinical data for cohort studies.

Conclusions

We derived model cohorts of quasi-healthy subjects, extracted from clinical data, by properly setting the extraction conditions. We then statistically compared the model cohorts with the control cohorts derived from a database of public checkup examinees in the same region. The distributions of clinical laboratory test values for both cohorts were similar. Although several problems remain to be solved, this method opens up new possibilities for using clinical data for cohort studies.

Acknowledgements

The authors thank the Kochi Health Research Institute for permission to use the Public Checkup DS. They would also like to thank Editage (www.editage.jp) for English language editing.

Author Contributions

SI and YK conceived and designed the experiments, contributed to the writing of the manuscript, agreed with manuscript results and conclusions, jointly developed the structure and arguments for the manuscript, and made critical revisions and approved the final version. SI analyzed the data and wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

Disclosure and Ethics

As a requirement for publication, the authors have provided to the publisher a signed confirmation of compliance with legal and ethical obligations, including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality, and (where applicable), protection of human and animal research subjects. The authors have read and confirmed their agreement with the *ICMJE* authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. The external blind peer reviewers report no conflicts of interest.

REFERENCES

1. Maslove DM, Rizk N, Lowe HJ. Computerized physician order entry in the critical care environment: a review of current literature. *J Intensive Care Med.* 2011;26:165–171.
2. Hillestad R, Bigelow J, Bower A, et al. Can electronic medical record systems transform health care? potential health benefits, savings, and costs. *Health Aff (Millwood).* 2005;24:1103–1117.
3. Spasic I, Sarafraz F, Keane JA, et al. Medication information extraction with linguistic pattern matching and semantic rules. *J Am Med Inform Assoc.* 2010;17:532–535.
4. Ohno-Machado L. Computer-based safety surveillance and patient-centered health records. *J Am Med Inform Assoc.* 2012;19:1.
5. de Bruijn W, Ibáñez C, Frisk P, et al. Introduction and utilization of high priced HCV medicines across Europe; implications for the future. *Front Pharmacol.* 2016;7:197.
6. Sugita S, Chikuda H, Kadono Y, et al. Clinical characteristics of rheumatoid arthritis patients undergoing cervical spine surgery: an analysis of National Database of Rheumatic Diseases in Japan. *BMC Musculoskelet Disord.* 2014;15:203.
7. Rothman KJ. *Epidemiology: An Introduction.* 2nd ed. Oxford, UK: Oxford University Press; 2012: 79–80.
8. Hoffmann RG. Statistics in the practice of medicine. *JAMA.* 1963;185: 864–873.
9. Ceriotti F, Boyd JC, Klein G, et al; IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). Reference intervals for serum creatinine concentrations: assessment of available data for global application. *Clin Chem.* 2008;54:559–566.
10. Ichihara K, Boyd JC; IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). An appraisal of statistical procedures used in derivation of reference intervals. *Clin Chem Lab Med.* 2010;48:1537–1551.
11. Smith HW. Plato and Clementine. *Bull N Y Acad Med.* 1947;23:352–377.
12. Nosanchuk JS, Gottmann AW. CUMS and delta checks. A systematic approach to quality control. *Am J Clin Pathol.* 1974;62:707–712.
13. Ichihara K, Kawai T. An iterative method for improved estimation of the mean of peer-group distributions in proficiency testing. *Clin Chem Lab Med.* 2005;43:412–421.
14. Dronina Y, Yoon YM, Sakamaki H, et al. Health system development and performance in Korea and Japan: a comparative study of 2000-2013. *J Lifestyle Med.* 2016;6:16–26.
15. Sasaki T, Izawa M, Okada Y. Current trends in health insurance systems: OECD countries vs. Japan. *Neurol Med Chir (Tokyo).* 2015;55:267–275.
16. OECD Health at a Glance 2013. http://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2013_health_glance-2013-en. Accessed August 14, 2017.
17. Tamura T, Kimura Y. Specific health checkups in Japan: the present situation analyzed using 5-year statistics and the future. *BMEL (Tokyo).* 2015;5:22–28.
18. Statistics and Information Department. Report on Regional Public Health Services and Health Promotion Services. http://www.mhlw.go.jp/english/database/db-hss/dl/rrphshps_2012.pdf. Accessed August 14, 2017.
19. Yamamoto K, Ogura H, Furutani H, et al. Efficacy of a computerized system applied to central operating theatre for medical records collection. *Med Inform (Lond).* 1986;11:329–338.
20. Ogura H, Sagara E, Yamamoto K, et al. Analysis of the online order entry process in an integrated hospital information system. *Comput Biol Med.* 1985;15:381–393.
21. Tominaga M, Makino H, Yoshino G, et al. Japanese standard reference material JDS Lot 2 for haemoglobin A1c. II: present state of standardization of haemoglobin A1c in Japan using the new reference material in routine clinical assays. *Ann Clin Biochem.* 2005;42:47–50.
22. Kumari NS, Rosario SB, Damodara Gowda KM. Altered liver function and the status of calcium in postmenopausal women in and around Mangalore. *Al Ameen J Med Sci.* 2010;3:115–119.
23. Aragon G, Younossi ZM. When and how to evaluate mildly elevated liver enzymes in apparently healthy patients. *Cleve Clin J Med.* 2010;77:195–204.
24. Djahanbakhch O, Ezzati M, Zosmer A. Reproductive ageing in women. *J Pathol.* 2007;211:219–231.
25. Portal Site of Official Statistics of Japan. Volume of sales (consumption) by prefectures 2007. https://www.e-stat.go.jp/SG1/estat/GL08020103.do?_toGL08020103_&listID=000001106932&disp=Other&requestSender=estat. Accessed August 14, 2017.